Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6802–6807
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

MMDAG: Multimodal Directed Acyclic Graph Network for Emotion Recognition in Conversation

Shuo Xu[1,2], Yuxiang Jia[1*], Changyong Niu[1], Hongying Zan[1]
1. School of Computer and Artificial Intelligence, Zhengzhou University, China
2. Zhengzhou Zoneyet Technology Co., Ltd., Zhengzhou, China
zzunlpxs1306@163.com, {ieyxjia, iecyniu, iehyzan}@zzu.edu.cn

Abstract

Emotion recognition in conversation is important for an empathetic dialogue system to understand the user's emotion and then generate appropriate emotional responses. However, most previous researches focus on modeling conversational contexts primarily based on the textual modality or simply utilizing multimodal information through feature concatenation. In order to exploit multimodal information and contextual information more effectively, we propose a multimodal directed acyclic graph (MMDAG) network by injecting information flows inside modality and across modalities into the DAG architecture. Experiments on IEMOCAP and MELD show that our model outperforms other state-of-the-art models. Comparative studies validate the effectiveness of the proposed modality fusion method.

Keywords: Multimodal, Directed Acyclic Graph Network, Emotion Recognition in Conversation

1. Introduction

Emotion can influence how a person thinks, feels and behaves, and it is an important part of human daily life. The Emotion Recognition in Conversation (ERC) task aims to identify the emotion of each utterance in a conversation, which draws increasing attention from researchers due to its potential applications in many domains, such as emotional conversation generation, sentiment analysis in social media (Chatterjee et al., 2019), psychoanalytic diagnostics, and understanding students' frustration in education (Poria et al., 2019b).

The rapid growth of conversation data on major social media is another drive of the popularity of the ERC task. Especially those open-sourced datasets, such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), and many others, greatly promote the research of ERC.

Researchers have proposed different models to utilize the contextual information of a conversation, such as DialogueRNN (Majumder et al., 2019) and DialogueGCN (Ghosal et al., 2020b). Basically, these models are usually classified into two categories: recurrence-based models and graph-based models. For recurrence-based models, the utterances are usually temporally encoded to consider distant utterances and sequential information while limited information from temporally nearest utterances makes it difficult to get a satisfactory performance. For the graph-based models, the information of the surrounding utterances within a certain window is concurrently gathered while distant utterances and sequential information is usually ignored. DAG-ERC (Shen et al., 2021b) integrates advantages of both recurrence-based model and graph-based model by representing a conversation as a directed acyclic graph (DAG) and proposes a DAG network to gather information from both neighboring utterances and remote utterances, which achieves state-of-the-art (SOTA) performance on several benchmarks. However, DAG-ERC only uses textual modality information while acoustic and visual information has also been proven useful for ERC.

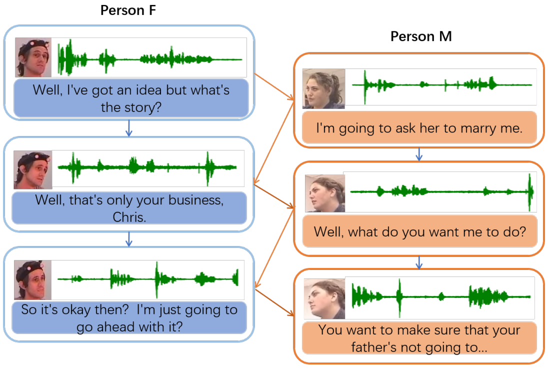
Person F
Well, I've got an idea but what's the story?
Well, that's only your business, Chris.
So it's okay then? I'm just going to go ahead with it?
Person M
I'm going to ask her to marry me.
Well, what do you want me to do?
You want to make sure that your father's not going to...

Figure 1: A directed acyclic graph of a conversation, with the brown edges representing the propagation of information between different speakers and the blue edges representing the propagation of information of the same speakers.

In order to model conversation contexts and utilize multimodal information more effectively, we extend DAG-ERC by proposing a multimodal DAG (MMDAG) to represent a conversation which can fuse textual, acoustic and visual features, as shown in figure 1. In MMDAG, information is only allowed to flow from previous utterances to the current utterance and the backward propagation from the current utterance to itself and its predecessors is not allowed. For each utterance, we have three nodes each denoting one

[*]Corresponding author

modality and allow information flows inside modality and across modalities.

Experimental results on the two benchmark multimodal dialogue datasets IEMOCAP and MELD show that MMDAG outperforms DAG-ERC, MMGCN (Hu et al., 2021b) and many other SOTA models. In addition, our proposed modality fusion method is more effective than comparative methods.

The rest of this paper is organized as follows. Section 2 summarizes related work about ERC and multimodal fusion. Section 3 presents the MMDAG model in detail. Section 4 shows experiments and analysis while section 5 draws conclusions.

2. Related Work

2.1. Emotion Recognition in Conversation

Contextual information is at the heart of the ERC task and thus many previous studies focus on modeling contextual information. There are primarily two categories of models, including recurrence-based models and graph-based models.

Recurrence-based models. BC-LSTM (Poria et al., 2017) uses an LSTM-based model to capture interaction history. CMN (Hazarika et al., 2018b) exploits internal speaker dependencies and uses speaker-dependent GRUs combined with conversation history information to model utterance contexts. However, CMN does not exploit the interaction information between speakers, a problem addressed by ICON (Hazarika et al., 2018a), which connects the historical utterances of two speakers with the same layer of GRUs. Similar to ICON and CMN, IANN (Yeh et al., 2019) employs a different memory for each speaker. DialogueRNN (Majumder et al., 2019) utilizes three GRUs to obtain contextual information and update speaker states.

Graph-based models. DialogueGCN (Ghosal et al., 2020b) treats each conversation as a graph and uses graph structure to model contexts. RGAT (Ishiwatari et al., 2020) adds positional encoding to DialogueGCN. KET (Zhong et al., 2019) utilizes hierarchical Transformers with external knowledge (Vaswani et al., 2017). DialogXL (Shen et al., 2021a) improves XLNet (Yang et al., 2019) with enhanced memory and dialogue-aware self-attention.

DAG-ERC (Shen et al., 2021b) encodes utterances with directed acyclic graphs and proposes a DAG neural network which can combine advantages of both recurrence-based model and graph-based model. COSMIC (Ghosal et al., 2020a) uses commonsense knowledge to learn interlocutor interaction to solve the task of ERC. TODKAT (Zhu et al., 2021) models the complex patterns of transformation between the topic, relevant commonsense knowledge and affective states behind a conversation. DialgueCRN (Hu et al., 2021a) proposes a new contextual reasoning network to comprehensively understand conversational contexts from a cognitive perspective.

2.2. Multimodal Fusion

ERC models such as CMN, ICON exploit multimodal information by concatenating three-modal features without exploring the interaction between the modalities. The TFN (Zadeh et al., 2017) model is a tensor fusion network that obtains a new tensor representation by computing the outer product between the unimodal representations. LMF (Liu et al., 2018) uses a low-rank multimodal fusion approach to decompose the weight tensor, reducing the computational complexity of the tensor-based approach. MFN (Zadeh et al., 2018) aligns features from different modalities well by fusing multi-view information. MulT (Tsai et al., 2019) is a cross-modal Transformer that learns attention between two-modal features, thus enabling implicit enhancement of the target modality without aligned data. MMGCN (Hu et al., 2021b) proposes a multimodal fusion graph convolutional network for ERC and discusses the impact of fusion methods of various modalities.

3. MMDAG for ERC

A conversation can be defined as a sequence of utterances $\{u_1, u_2, ..., u_N\}$, where N is the number of utterances in the conversation. Each utterance involves three sources of data corresponding to three modalities, including acoustic, visual and textual modalities, as shown in equation 1.

$$u_i = \{u_i^a, u_i^v, u_i^t\} \tag{1}$$

where u_i^a, u_i^v, u_i^t denote the raw feature representation of u_i from acoustic, visual and textual modalities respectively. The goal of ERC task aims to predict the emotion label for each utterance in the conversation based on the available information from three modalities.

In order to utilize multimodal information and model conversational contexts effectively, based on DAG-ERC (Shen et al., 2021b), we propose MMDAG model for ERC. The overall architecture of MMDAG is shown in figure 2. In the MMDAG layer, a box of the same color represents an utterance of the same speaker, solid lines represent the edges of local information, dash lines denote the edges of remote information, and the color of a line is the same with the end box.

To cope with multimodal ERC, we make the following extensions to DAG-ERC:

(1) For each node u_i in DAG-ERC, we create three nodes u_i^a, u_i^v, u_i^t. Thus the number of nodes in MMDAG will be tripled.

(2) For different utterances, we constraint that links can only be established between nodes of the same modality. This constraint is reasonable and makes the enlarged graph simple.

(3) For the same utterance, we only allow information flows from acoustic node and visual node to textual node. This constraint keeps the graph acyclic. Another reason is the experimental evidence that the tex-

tual modality is much more important than other two modalities.

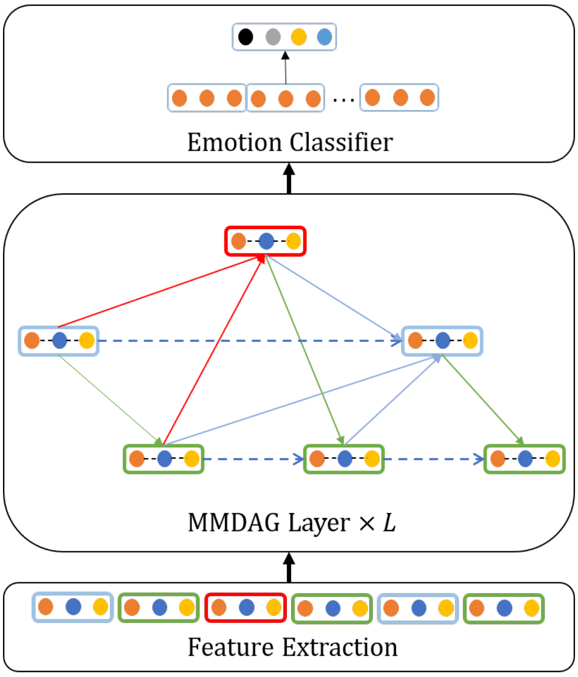

Figure 2: The overall architecture of MMDAG for ERC.

The textual raw features are extracted using RoBERTa-Large (Liu et al., 2019). We follow MMGCN (Hu et al., 2021b) to extract the acoustic raw features with the OpenSmile toolkit while extract the visual facial expression features with a pre-trained DenseNet. The representation of extracted features goes through L layers of MMDAG network and the outputs for each utterance u_i are three hidden states corresponding to the three modalities as shown in equation 2 to 4. The hidden state of each modality is the concatenation of all hidden states of that modality at all MMDAG layers. The final representation of utterance u_i is the summation of hidden states of three modalities, as shown in equation 5.

$$H_{ai} = ||_{l=0}^{L} H_{ai}^{l} \quad (2)$$

$$H_{vi} = ||_{l=0}^{L} H_{vi}^{l} \quad (3)$$

$$H_{ti} = ||_{l=0}^{L} H_{ti}^{l} \quad (4)$$

$$H_i = H_{ai} + H_{vi} + H_{ti} \quad (5)$$

Finally, we pass H_i through a feed-forward neural network to get the predicted emotion. For the training of MMDAG, we use the standard cross-entropy loss as the objective function.

4. Experiments and Analysis

4.1. Experimental Settings

Two public multimodal conversation datasets, IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a), are used for experiments. The former is a two-party scenario while the latter is a multi-party scenario. IEMOCAP contains six emotions, including *happy*, *sad*, *neutral*, *angry*, *excited*, and *frustrated*. MELD contains seven emotions, including *happy*, *sad*, *neutral*, *angry*, *surprised*, *disgusted*, and *fearful*. As previous studies, we divide the datasets into training/validation and test sets with a ratio of approximately 8:2. Detailed statistics of data distribution are shown in table 1.

Dataset	#Conversation		#Utterance	
	train	test	train	test
IEMOCAP	120	31	5810	1623
MELD	1152	280	11098	2610

Table 1: Statistics of data distribution

Hyper-parameters	IEMOCAP	MELD
Local window size *w*	1	
#DAG Layers	4	2
Dropout rate	0.4	
Learning rate	0.0005	0.0004
Size of hidden vectors	300	
RoBERTa feature size	1024	

Table 2: Hyper-parameter settings

The hyper-parameters of MMDAG model on IEMOCAP and MELD are set as shown in table 2. Both of values of the local window size *w* equal 1. The number of DAG layers are 4 and 2 respectively. Both of the dropout rates are 0.4. The learning rates are 0.0005 and 0.0002 respectively. Both of the sizes of all hidden vectors are 300, and the RoBERTa extracted feature sizes are 1024. The training and testing processes run on a single RTX2080Ti GPU.

4.2. Comparison with Baseline Models

Baseline models include BC-LSTM (Poria et al., 2017), ICON (Hazarika et al., 2018a), DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2020b), MMGCN (Hu et al., 2021b), DAG-ERC (Shen et al., 2021b) and two variant models of MMGCN. Among those models, DialogueGCN and DAG-ERC are based solely on textual modality while others are based on multimodality. Weighted-average F1-score is used as the evaluation metric.

The ERC results of different models on IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a) are shown in table 3. DAG-ERC model outperforms all other baseline models on both datasets. When incorporating multimodal information, our MMDAG model further improves F1-score over DAG-ERC by 0.98% and 0.06% respectively on the two benchmarks.

When RoBERTa is used for text feature extraction in MMGCN, performance of MMGCN-RoBERTa gets improved. MMGCN-RoBERTa' has the same modules of multimodal feature extraction and emotion

Model/Dataset	IEMOCAP	MELD
BC-LSTM	54.95	56.8
ICON	58.54	—
DialogueRNN	64.58	57.11
DialogueGCN	65.04	58.23
MMGCN	66.22	58.62
MMGCN-RoBERTa	67.58	—
MMGCN-RoBERTa’	67.18	—
DAG-ERC	68.03	63.65
MMDAG	**69.01**	**63.71**

Table 3: F1-scores of different models on IEMOCAP and MELD

Modality	IEMOCAP	MELD
a	52.59	36.70
v	31.72	30.36
t	67.45	63.20
a-t	68.79	63.58
v-t	67.78	63.49
a-v-t	**69.01**	**63.71**

Table 4: Performance of MMDAG under different modality settings

classification but different middle network layers with MMDAG. The better performance of MMDAG over MMGCN-RoBERTa’ shows the superiority of graph structure of DAG over GCN on ERC task.

4.3. Modality Settings

Performance of MMDAG under different modality settings is shown in table 4. For single modal settings, textual modality model works the best, followed by acoustic modality model and visual modality model. When fusing acoustic and visual modality with textual modality, performance gets improved. The best performance is achieved by textual and acoustic fusing model, with 1.56% and 0.51% improvement of F1-score on IEMOCAP and MELD respectively over textual modality model.

4.4. Modality Fusion Methods

In MMDAG, we build a DAG with nodes of different modalities and allow information flows between nodes of the same modality and across modalities. To verify the effectiveness of this modality fusion way, we make comparisons with other modality fusion methods, including early fusion, late fusion, MFN (Zadeh et al., 2018) and MulT (Tsai et al., 2019).

As for the early fusion method, multimodal features are directly concatenated and then fed into the DAG. As for the late fusion method, features from different modalities are fed into separate DAGs and then the outputs are concatenated for emotion classification. In MulT fusion method, multimodal features are input to the MulT network for fusion before being fed into the DAG. Similarly, in MFN fusion method, features from different modalities are input to the MFN network for fusion before being fed into the DAG.

Model/Dataset	IEMOCAP	MELD
MulT-DAG	67.66	63.48
MFN-DAG	68.04	63.66
Early-fusion-DAG	67.78	63.36
Late-fusion-DAG	68.01	63.52
MMDAG	**69.01**	**63.71**

Table 5: Comparison of different modality fusion methods

The ERC results with different modality fusion methods on IEMOCAP and MELD are shown in table 5. MMDAG performs the best, which indicates that the proposed modality fusion method are more effective than other modality fusion methods.

4.5. Results of Different Emotions

Recognition results of MMDAG on IEMOCAP with different emotions are shown in figure 3. The F1-score for emotion *happy* is the lowest may be partly due to the lowest portion of data of this emotion. The F1-score of emotion *sad* is over 80% though it only accounts for 15% of all utterances.

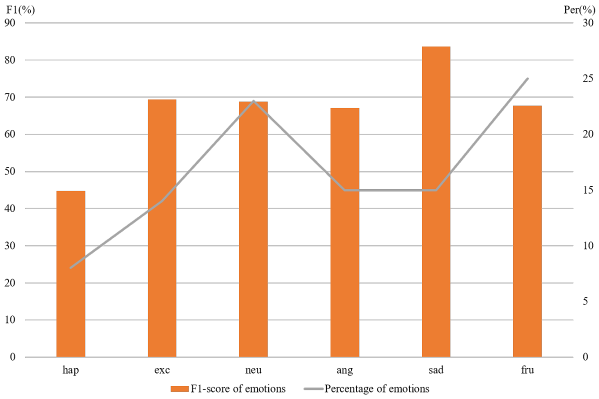

Figure 3: F1-score of MMDAG on IEMOCAP for different emotions in histogram, and percentage of data in line chart.

In order to investigate feature words of each emotion, we compute the tf-idf value of each word to each emotion, and sort words according to the descending order of the tf-idf value. The top 10 feature words for each emotion in IEMOCAP are listed in table 6. We can see some negative words in negative emotions like *angry* and *sad*.

5. Conclusions

In this paper, we propose a multimodal directed acyclic graph neural network model for emotion recognition

Emotions	Feature words
happy	beginning, wondering, white, climb, fortune, what'll, forgive, camping, i, you
excited	swimming, coast, hairs, mindedness, urgency, pure, urge, swallows, celebration, sex
neutral	taken, amount, seats, available, ahead, sir, you, craig's, time's, cash
angry	beast, vile, tempered, evil, insulting, listening, wicked, sadistic, bully, cruel
sad	ashamed, pictures, toss, killed, selfish, included, changed, overseas, instance, raining
frustrated	believes, pinpoint, lady, identification, resumes, breathe, snap, showed, northwest, diagram

Table 6: Top 10 feature words of each emotion in IEMOCAP

in conversation by extending DAG-ERC with modality fusion method. In MMDAG, information flows between nodes of the same modality and across modalities are exploited. Extensive experiments are carried out on benchmarks IEMOCAP and MELD. Experimental results show that MMDAG gets significant improvement over DAG-ERC and outperforms MMGCN and other baseline models by a large margin. Acoustic and visual information can enhance performance of textual ERC model. Our modality fusion method is more effective than other comparative modality fusion methods. In addition, we analyze the performance of MMDAG on different emotions .

6. Acknowledgements

This work was supported in part by National Key Research and Development Project (Grant No. 2017YFB1002101), Major Program of National Social Science Foundation of China (Grant No. 17ZDA138, 18ZDA295), National Natural Science Foundation of China (Grant No. 62006211), and China Postdoctoral Science Foundation (Grant No. 2019TQ0286, 2020M682349).

7. Bibliographical References

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., and Poria, S. (2020a). Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2470–2481.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2020b). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., and Zimmermann, R. (2018a). Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., and Zimmermann, R. (2018b). Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132.

Hu, D., Wei, L., and Huai, X. (2021a). Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Hu, J., Liu, Y., Zhao, J., and Jin, Q. (2021b). Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675.

Ishiwatari, T., Yasuda, Y., Miyazaki, T., and Goto, J. (2020). Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A. B., and Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019a). Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019b). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Shen, W., Chen, J., Quan, X., and Xie, Z. (2021a). Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Shen, W., Wu, S., Yang, Y., and Quan, X. (2021b). Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yeh, S.-L., Lin, Y.-S., and Lee, C.-C. (2019). An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689. IEEE.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., and Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhong, P., Wang, D., and Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

Zhu, L., Pergola, G., Gui, L., Zhou, D., and He, Y. (2021). Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.